ANALYSIS OF LOCAL EXPERTS IN SOCIAL MEDIA

A Thesis

by

SINDHUJA VENKATESH

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	James Caverlee
Committee Members,	Thomas R. Ioerger
	Daniel W. Goldberg
Head of Department,	Dilma Da Silva

December 2014

Major Subject: Computer Engineering

ABSTRACT

Recent popular social services (e.g., Foursquare, Twitter, Instagram) are creating a comprehensive geo-social overlay of the planet through geo-located posts, images, and other user-generated content. These public, voluntarily shared footprints provide a potentially rich source for uncovering the landscape of users' interests and topical expertise, which has important implications for social search engines, recommender systems, and other geo and socially-aware applications. This thesis presents the first large-scale investigation of local interests and expertise through an analysis of a unique 13 million user geo-coded list dataset sampled from Twitter. Twitter lists encode a "known for" relationship between a labeler and a labelee. In the small, these lists are helpful for individual users to organize friends or contacts. In the aggregate, however, these lists reveal global patterns of interest and expertise. Concretely, this thesis presents a qualitative and quantitative analysis on the relationships between user locations, interests, and topic expertise as revealed through these Twitter lists. Through thorough analysis this thesis examines the (i) impact of geo-location on topic expertise and users' topic interests in Twitter; (ii) the degree of "locality" of topics; and (iii) the concentration and dispersion of expertise.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Prof. James Caverlee, for his continued guidance for the work I have done so far. He has been a great source of inspiration and continues to motivate me to help me achieve my research objectives. I would like to express my sincere gratitude to my thesis committee members, Prof. Thomas Ioerger and Prof. Daniel Goldberg, for their advice, encouraging words and feedback on various milestones of this thesis. I would also like to thank Zhiyuan Cheng, Haokai Lu and my other colleagues from infolab for their support, technical advice and feedback during the various phases of my work. Lastly I would like to thank my friends Shiva Dhandapani and Srinivas Venkatasubramanian, Industrial Engineering graduates at Texas A&M, for their valuable technical inputs, constant motivation and support.

NOMENCLATURE

Tech	Technology
Celebs	Celebrities
SF	San Francisco
NY	New York

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

With the emergence of social media, more and more users have started to share their geolocation information online. For example, many social sites – including Facebook, Instagram, and Twitter – give users the option to provide their location information, ranging from a coarse "home location" to fine-grained GPS-tagged social media posts. Confirming this trend, a recent Pew Research Center report finds that location is now an increasingly central part of the social media experience [41]. Unlike proprietary location-based data (e.g., query logs, cell phone call records, point-of-sale data), these *geo-social* signals are inherently voluntary and public. These voluntary, publicly-shared signals provide the basis for new investigations into (i) the dynamics of human behavior and pulsation of social life from local to global levels; (ii) the dynamics of how ideas spread and how people can organize for societal impact; and (iii) the development of new geo-social information systems that leverage these global-scale geospatial footprints for real-world impact. Already, we have seen much research along these three aspects across multiple areas, including in data mining and machine learning [3, 7, 13, 16], in geographic information systems [9, 20, 32, 42, 45] and in web search and information retrieval [38, 41].

In this thesis, we are interested to explore geo-social signals as a potentially rich source to uncover the landscape of users' interests and topical expertise. Why do we care about geo-social signals for user interest and expertise? By providing a new perspective on user interests, search engines and social media sites can augment the discoverability of their own content, develop new recommender systems that explicitly leverage these geo-social patterns. There is evidence to show that when seeking an expert, users consider both the relevance of the person to the topic and

the network topology, for example their social distance to the expert [34]. Recently, there has been an increasing focus on geo-marketing and geo-targeted advertising – providing users with custom content that encompasses both topical relevance and geographical relevance. This goes to show that understanding geo-social signals would greatly enhance research systems to understand and meet users' need for information.

Additionally, sociologists and communications researchers have long pondered over the interplay of people's location, interactions and social ties. Each individual is tightly embedded in one's social structure and this social environment and geography play an important role in shaping the nature of people and information that one has access to. Over the years, many researchers have noted an inverse relationship between distance and the likelihood of friendship. Apart from likelihood of friendship, the density and spatial arrangement also is expected to have an impact on the size and frequency of interaction among social ties [33].

And in one important direction, these geo-social signals can provide a window into *local experts*. Local experts bring specialized knowledge about a particular location and can provide insights that are typically unavailable to more general topic experts. For example, a "foodie" local expert is someone who is knowledgeable about the local food scene, and may be able to answer local information needs like: what's the best barbecue in town? Which restaurants locally source their vegetables? Which pubs are good for hearing new bands? Similarly, a local "techie" expert could be a conduit to connecting with local entrepreneurs, identifying tech-oriented neighborhood hangouts, and recommending local talent (e.g., do you know any good, available web developers?). Indeed, a recent Yahoo! Research survey found that 43% of participants would like to directly contact local experts for advice and recommendations (in the context of online review systems like Yelp), while 39% would not mind be-

ing contacted by others [1]. And yet, there has been little research on identifying or analyzing local expertise, mainly due to the lack of large-scale publicly-available signals.

Figure 1.1: Example of a twitter list.

Towards bridging this gap, this thesis presents the first large-scale investigation of local interests and expertise through an analysis of a 13 million user-labeled dataset sampled from Twitter. Apart from its well-known "follow" feature, Twitter provides another feature called *lists* as a way to stay connected and network effectively. A Twitter list is a curated group of Twitter users. Twitter allows users to create their own lists or subscribe to lists created by others. Viewing a list timeline will show the users a stream of Tweets from only the users on that list, an example of which is shown in Figure 1.1. Thus it is an effective way of organizing one's Twitter feed into easily viewable categories. Lists are essentially user labeled topics, with each

list consisting of various Twitter users that the user perceives as belonging to that topic.

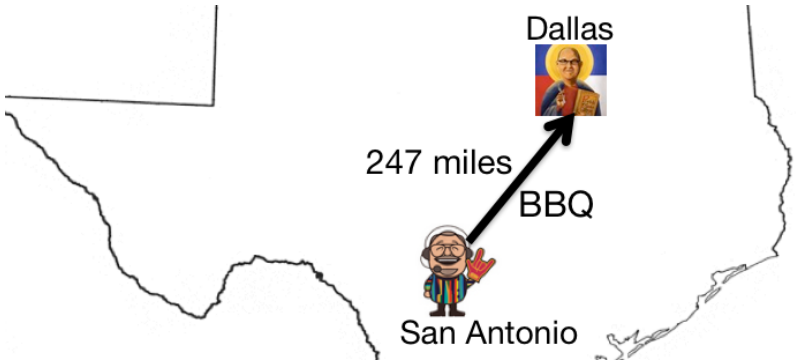
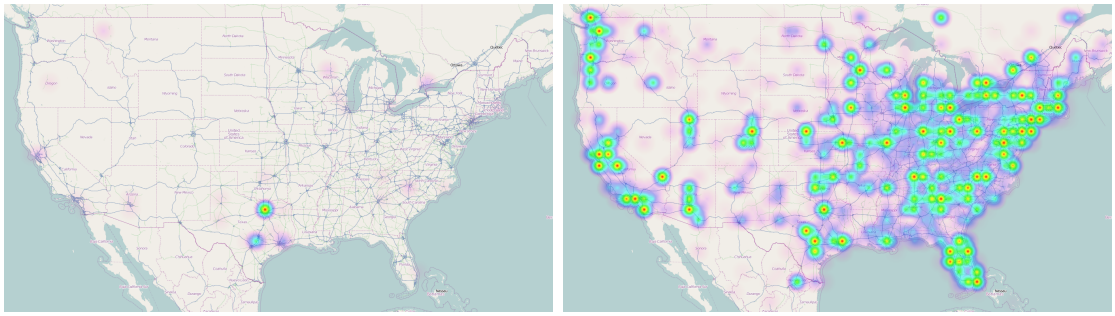


Figure 1.2: @jerry lists @BBQsnob with label 'bbq'

Thus lists are a special capability to capture a *known-for* relationship between two users. Typically, one user can add another user to a list with a particular annotation. In Figure 1.2, Twitter user @jerry has added another Twitter user @BBQsnob to a list labeled as `bbq`. In the small, these lists are helpful for individual users to organize friends or contacts. In the aggregate, however, these lists may reveal more global patterns, which in some cases may be interpreted as expertise. This investigation has not only the 13 million lists, but also the location of each user. Through this fine-grained geo-social perspective, we can study the interplay between location, interest, and topic expertise. As an example, we can see in Figure 1.3 the distribution of list labelers for two Twitter users: (a) @BBQsnob; and (b) @JimmyFallon. We can see that @JimmyFallon attracts a large following from across the country, whereas @BBQsnob is very popular, but primarily only in Texas.



(a) @BBQsnob (b) @JimmyFallon

Figure 1.3: Heatmap of the location of list labelers

Now the question arises – can we leverage these Twitter crowdsourced labels to begin an investigation of local interests and expertise? In this thesis, I use Twitter lists to conduct a thorough investigation of geographical impact on interests and expertise. Concretely, I perform a qualitative and quantitative analysis on the relationships between user locations, interests, and topic expertise through an investigation of Twitter lists. Specifically, this thesis addresses the following questions:

- Does geolocation play a role in topic interest and expertise?
- For one topic, do the geospatial footprint of interests and expertise vary?
- Are some topics inherently more local than others?
- Is level of expertise uniformly spread out among all the experts in a topic?
- Does expertise affect the locality of an expert?

The remainder of the thesis is organized as follows. Chapter 2 presents a literature survey. Chapter 3 presents the data-driven analysis of local expertise and the main findings of the thesis. Chapter 4 concludes with some final thoughts and

opportunities for future work that builds on the mainly observational nature of this thesis.

2. RELATED WORK

My work in the thesis here builds on three lines of research: study of Twitter lists, local experts, and geo-spatial analysis of social media.

2.1 Twitter Lists

There have been several papers studying the general properties of Twitter as a social network and in analyzing information diffusion over this network [31]. Research into Twitter lists is still in its nascent stages. Kim et. al [29] perform an analysis of Twitter lists as a potential source for discovering latent characteristics of users. They conducted a study using Twitter lists to infer characteristics and interests of the users in those lists. They showed that by using the tweets of all the users in the list, they could discover the characteristics and interests of the users in that list, even if the users as individuals do not tweet about that topic. Their experiments confirmed that the user interests, as found by their system using Twitter lists, reflect the interests that are perceived by the human subjects in their user survey. Their study showed that their approach yielded good agreements between human decision and list tags even for the words that are not in the user's timeline. Based on their results, they proposed a list of potential research on Twitter lists: i) Expert Recommendation System: A Twitter list may consist of users with expertise on same topic. They suggest that their work could be extended to discover experts on Twitter. ii) Information Source: Many users are sharing up-to-date news and events on Twitter. They claim that as most geographic lists are composed of people who live in or know well about the locations, tweets in these lists serve as local news. This provides motivation to analyze Twitter lists to understand topic expertise and local expertise. Yamaguchi et. al [46] address the problem of tagging users in Twit-

ter using Twitter Lists. They argue that extracting tags from tweets needs a lot of preprocessing and tweets do not always contain keywords which represents the user topic (Topic of expertise). Instead they exploit Twitter lists to tag users based on the membership of lists with the tag name.

2.2 Local Experts on Social Media

There has been a considerable amount of work done to identify topic experts among Twitter users. Bernstein et. al designed Collabio [4], a tagging-based Facebook game that encouraged users to tag people in their networks. The metadata collected by the game about users was intended to be used to find experts in social media. Weng et. al [44], proposed a ranking similar to Page-Rank, called TwitterRank, that uses the information from Twitter social graph and information from tweets to identify experts in specific topics. Pal et. al [36] used a set of 15 features extracted from the Twitter graph and tweets posted by the users to estimate their expertness in topics. Ghosh et. al. [21] devised a system called Cognos which used Twitter Lists feature, which are user-curated lists of people, to identify topic experts and claim to perform better than graph and tweet feature based expert finding systems. Aardvark.com [24] a commercial social expert finder, which tries to address the challenge of determining the right person for a person's information need. They studied how factors like trust due to intimacy, user's social graph, etc. influenced a person's information need and the quality of answers.

2.3 Geo-spatial Analysis of Social Media

The emergence of location-based social networks like Foursquare, Gowalla, and Google Latitude has motivated large-scale geo-spatial analysis [27, 39, 35, 8]. Some of the earliest research related to geo-spatial analysis of web content were based on mining geography specific content for search engines [14]. In [2] the authors analyzed

search queries to understand the spatial distribution of queries and understand their geographical centers. In Ghosh et. al [20], the authors chose obesity as a test theme to demonstrate the effectiveness of topic modeling using Latent Dirichlet Allocation (LDA) and spatial analysis using Geographic Information System (GIS). Facebook researchers have provided a comprehensive analysis of the distance between Facebook users, leading to new insights into how social networks are impacted by geography [3]. The LiveHoods [10] project has shown how to identify "living neighborhoods" based on the revealed locations and movements of social media users. On Twitter, geo-spatial analysis has focused on inferring geographic information from tweets like predicting user locations from tweets [7] and spatial modeling to geolocate objects [12]. Adam et. al [37] built a system called Flap that claimed to i) reconstruct the entire friendship graph with high accuracy even when no edges were given; ii) inferred people's fine-grained location, even when they keep their data private and only friends' location was accessible. They used a combination of multiple disparate features, based on text, location and topology of the underlying friendship graph. Researchers have also analyzed Youtube videos for geo-spatial properties and observed the highly-local nature of video views [5]. Through projecting of users' social network structure onto space, authors in [11] attempt to discover knowledge on the ties among distributed clusters of communities in the real world. Location recommendation systems with emphasis on spatial nature in the past human behavior and information about the user social interaction with other users have shown to outperform traditional recommendation systems [43]. Probabilistic topic models approach was used by authors in [17] to extract urban patterns from location based social network data. They observed that the extracted patterns can identify hotspots in the city, and recognize a number of major crowd behaviors that recur over time and space in the urban scenario. Through such research works, it is apparent that geo-

spatial signals in social media have strong ties to real world user behavior and thus studying them would greatly help us address the problems and needs of users.

My work is closely related to Cheng et all [6], that encompasses the above three research areas. They use data from Twitter lists, tweets and the social graph of the user to propose a geo-spatial approach to finding local experts on Twitter. They proposed a local expertise framework that integrates both users' topical expertise and their local authority. They estimated a user's local authority through spatial proximity expertise approach using geo-tagged Twitter lists. They estimate a user's topical expertise based on expertise propagation. Through their initial analysis, they concluded that certain topic are inherently more local and that identifying local experts in topics that are inherently more local could be easier than identifying local experts in other topics.

In my work I consider user's relation to a topic on two semantics – 1) the user is an *expert* in the topic and 2) the user is *interested* in the topic. The comparison between the two semantics was studied in [22]. The authors perform an extensive study that explores the use of social media to infer expertise within a large global organization. They examine eight different social media applications and evaluate their results through a large user survey. In their work, they use self-identified user ratings for expertise evaluation, while in this thesis, I use topic labels assigned by other users as a signal for expertise.

Using this as a precursor, this thesis performs a comprehensive analysis of the impact of geolocation and geographical distances on expert finding in different topics.

3. DATA-DRIVEN ANALYSIS OF LOCAL EXPERTISE

In this chapter, I present my data-driven analysis of geo-located Twitter lists. Recall that lists are essentially user-labelled topics, with each list consisting of various Twitter users that the user perceives as belonging to that topic. In many cases, list names represent topics that the list members are strongly associated with. For example, when a user wants to group some user accounts who often post tweet about technology, he/she may create a list named `technology` or `tech` consisting of those accounts. The maximum length of list names is 25 characters. Most of list names are sequences of terms connected by delimiters. This feature partly motivates the use of list names as topics to tag the users in the list.

The lists bring out two interesting features: i) interests; and ii) expertise. A Twitter list is an effective way of organizing one's Twitter feed into easily viewable categories. When a user creates a list and tags it with a name, it is a signal that identifies the user's interest in that topic/category as the user wishes to view tweets from users present in the list. When a user is present in a list tagged with a topic, it can be considered as a signal that the user is 'known for' or 'belongs to' that topic/category. The larger the number of lists a user appears in, the more popular he/she is in that topic. Inherently, this indicates the user's 'expertise' in that topic as perceived by others. I study these two features brought out by lists. Concretely, the study in this chapter is organized as follows:

- **Data and Setup** [Section 3.1]: First, I detail the Twitter data and present some general descriptive characteristics of the users and the tags that they employ.
- **Localness of Experts** [Section 3.2]: Next I investigate the properties of ex-

pertise manifested by the dataset. Here I address the questions 1) Is the effect of geolocation the same for a topic for labelers and experts? 2) Does expertise distribution is consistent among all experts in a topic? How does it vary across topics?

- **Localness of Topics** [Section 3.3]: Then, I look into the localness of topics, themselves. Are some topics inherently more local than others? How do the spheres of influnce of experts from different topics vary?Twitter
- **Local vs Global: Measuring Focus, Entropy, and Spread** [Section 3.3.1]: Then, I capture the localness of topics in a quantitative sense. I study the localness of a topic at a micro-level by quantifying the localness of individual experts.
- **Popularity Vs Localness** [Section 3.4]: Next, I study how popularity of an expert affects these measures. Through this analysis, I address the questions - Do popular experts from different topics exhibit the same properties? What is the relation between the popularity of an expert and his/her sphere of influence?
- **How Does Topic Localness Vary Across Locations?** [Section 3.5]: Finally, I drill deeper into locations and measure localness at cities level. Here I look at the variance of localness for a topic across different cities within the United States. Given a topic with high localness signal, is the geolocation effect the same across locations? Does popularity of a topic vary across cities? How does popularity affect localness of the topic?

3.1 Data and Setup

I sample 54 million Twitter user profiles, as well as 3 billion geo-tagged tweets. For each user, I seek to assign a home location; however it is widely observed that many

Twitter users reveal overly coarse or no location at all in the self-reported location field. Hence, I adopt a home finding method that relies on a user's geo-tagged tweets akin to a similar approved previously used for check-ins and geo-tagged images. First I group the user's locations where he posted his tweets into squares of one degree latitude by one degree longitude (covering about 4,000 square miles). Next I select the square containing the most geo-tagged tweets as the center, and select the eight neighboring squares to form a lattice. I divide the lattice into squares measuring 0.1 by 0.1 square degrees, and repeat the center and neighbor selection procedures. This process repeats until I arrive at squares of size 0.001 by 0.001 square degrees (covering about 0.004 square miles). Finally, I select the center of the square with the most geo-tagged tweets as the 'home' of the user. In total, I geo-locate about 24 million out of the 54 million users (about 45.1%) with fine-grained latitude-longitude coordinates. Out of the 24 million Twitter users, I sample 13 million lists that these users occur or that the users have created. This set consists of 14.7 million pairs of geo-location list relationships indicating a direct link from a list creator's location to a list member's location.

Considerable research has been done towards geo-coding socially generated data [19, 26, 23]. Location determination is not a trivial task. The simplest method is to consider the user declared location in profile information. Since it is the form of free text, it is often hard to geolocate correctly. High error rates, missing data and non-standardized text in profile locations led researchers to explore other manual coding methods. One option is to use only geo-coded tweets, which is either an exact location specified as a pair of latitude and longitude coordinates or an approximate location specified as a bounding box. Due to privacy risks, tweet geolocation is disabled by default and users must explicitly alter their account settings to enable it. Only a small portion of users publish geocoded tweets, and it is unlikely that they form a

representative sample of the broader universe of content (i.e. the division between geocoding and non-geocoding users is almost certainly biased by factors such as social-economic status, location, education, etc.) On a typical day only about 1.5% of the tweets are referenced with exact location. Around the same fraction of tweets have location information in the form of place indicators in textual form. This work uses the geographical data obtained from tweets locations to perform very coarse spatial analysis - at country and city levels only. Thus, the caveats associated with this method would not adversely affect the study.

Data Type	Total # of Records
Lists	12,882,292
Geo-Tagged List Relationships	14,763,767
Unique tags from lists names	230,073

Table 3.1: Geo-tagged twitter data

First, I analyze and report some preliminary characteristics of the dataset, including the users and the tags (labels) that they employ:

Users: To compute usage statistics, I investigate the geo-tagged users to obtain the distribution of lists created frequency that is shown in figure 3.1. The frequency of lists created is plotted on the X-axis and the number of users who created that many lists is plotted on the Y-axis. From this it can be observed that there is a steep downward trend in the plot. The majority of users create only a few lists, 10 lists or fewer. Very few users create many lists; it can be seen that only around 10 users have created 50 lists or more. This is encouraging since it means the labels are not

dominated by a handful of super-users, but rather, they reflect a wide crowdsourced perspective on what Twitter users are known for.

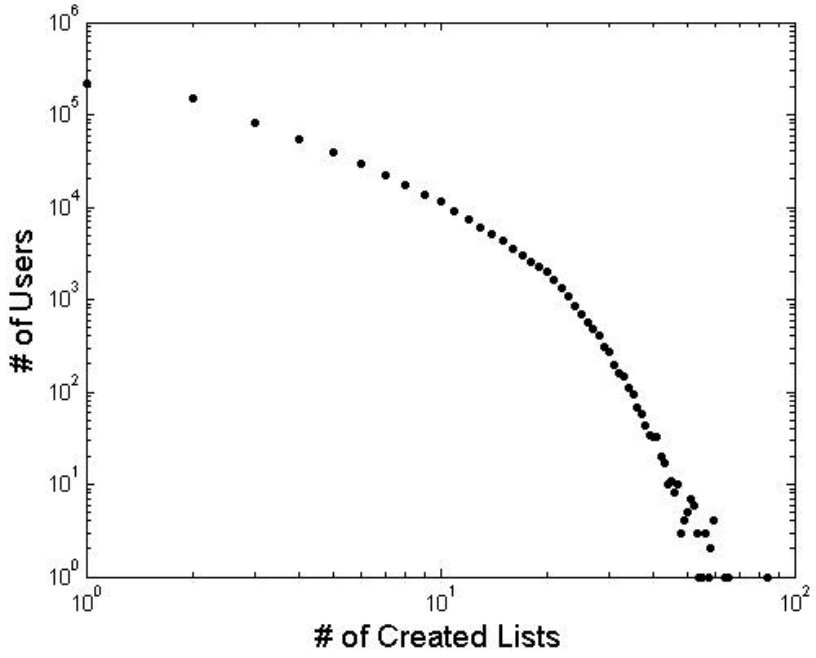


Figure 3.1: The distribution of the number of lists created per user

Figure 3.2 shows the distribution of inclusion frequency. The horizontal axis is the number of lists and the vertical axis is the number of users who are included in the corresponding number of lists. It can be seen that there is a peak around 13 lists – that is, the median number of lists a user appears on is 13. It can also be seen that there are some super-users who appear on 100s of lists (the lower righthand portion of the figure). This is encouraging since it means that many users are included on lists (and not just a handful of celebrities) and that most users belong to many lists (so that their expertise is reflected in the viewpoints of many labelers).

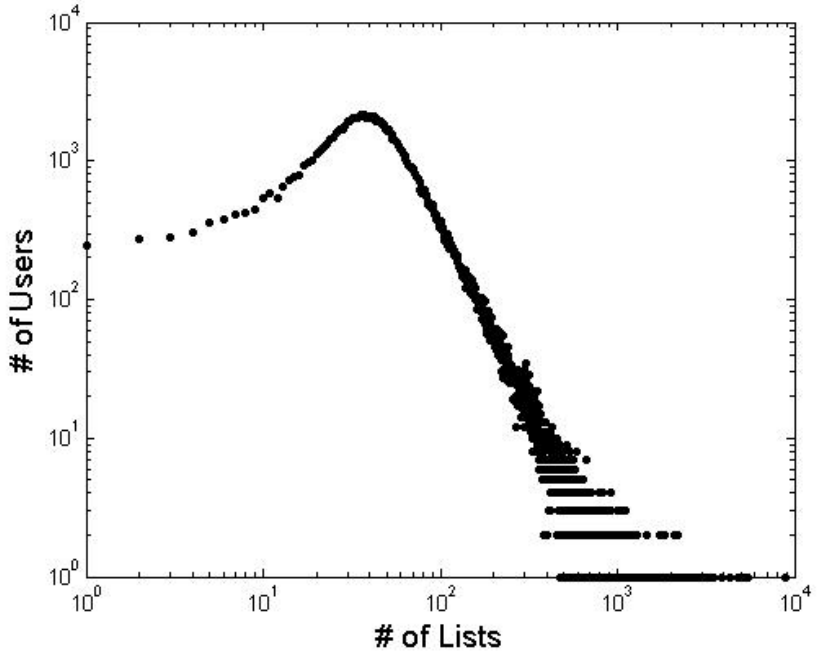


Figure 3.2: The distribution of frequency of list membership

Labels: Since the list names are user-generated, there is a lot of noise in the data thus requiring pre-processing of the names to extract tags from them. All the list names were converted to lower case, tokenized using punctuations as delimiters, Porter stemmer was used to stem the label names. The tag names thus obtained was used to tag the lists.

Figure 3.3 shows the distribution of tag occurrence frequency. The distribution follows a power law, meaning that small number of tags occur frequently from list names and most of the tags occur a few times. Hence, user selection of tags is highly concentrated. In terms of assessing user topics and expertise, this is encouraging since there is a common "language" for tagging as reflected in the dominance of certain tags. While there is some variability on any specific topic – e.g., the tags `tech`, `technology`, `techie` are all broadly related to the topic of *technology* – users

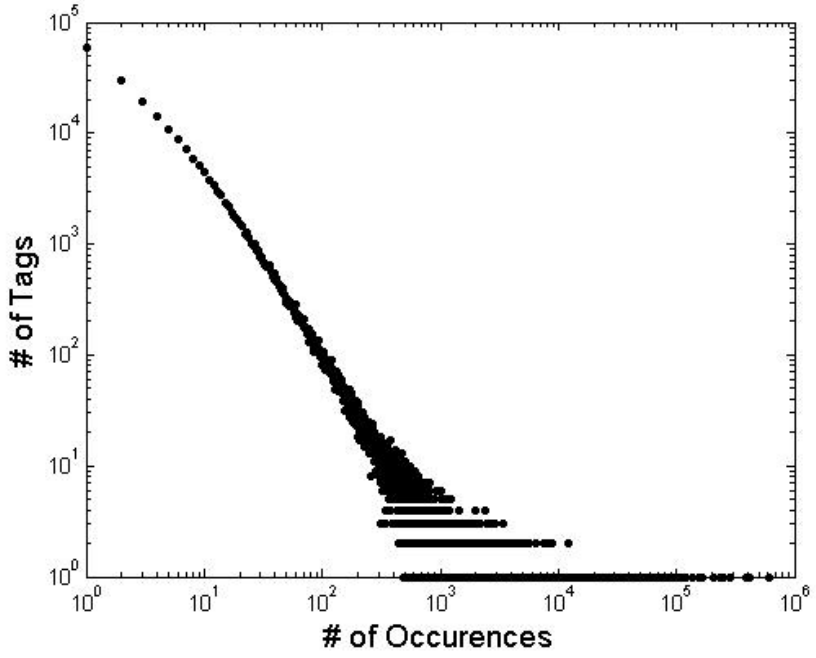


Figure 3.3: The distribution of tag occurrence frequency

tend to apply tags that are broadly used by others.

To further illustrate, Table 3.2 shows the top-ten most frequent tags. It can be seen that the tag `news` is the most popular, appearing in over 600,000 user-generated lists. As observed in [46], most of the tags extracted from list names are nouns. In general, there are typically four kinds of tags:

- Topic Tags, which describe user topics (e.g., `music` or `sports`)
- Property Tags, which show the property of users (e.g., `famous`, `politician`)
- Personal Tags, which make sense only from the labelers" personal viewpoint (e.g., `friend`, `conversation`)
- Nonsense Tags, which do not make any sense as semantically meaningful tags (e.g., `list`, `and`)

Rank	Tag	Count	Rank	Tag	Count
1	news	607,607	6	people	236,589
2	media	421,592	7	social	234,611
3	music	385,869	8	celebs	210,172
4	twibes	280,503	9	sports	206,646
5	tech	252,535	10	marketing	165,289

Table 3.2: Top ten most frequent tags

In this thesis, I focus on tags belonging to the first and second categories. The third and fourth categories are not relevant in this context and will not be considered. For in-depth analysis, I focus my investigation in the remainder on the following topics: news, media, tech, celebs, food, finance, politics, travel and sports. The topic was generated by combining tags that are similar in nature. Table 3.3 shows the tags for the topics considered. Also, I consider geo-locations that fall within the boundaries of United States of America.

Topic	Tag names grouped
news	news
media	media
tech	tech,technology,techies,techno,techie,techy
celebs	celeb,celebs,celebrity,celebrities,celebz
food	food, foods, foodie, foodies
finance	finance, finances
politics	politics, political, politica, politico, politicians, politician

Table 3.3: Topics derived from tag names

3.2 Localness of Experts

In this section, I begin my investigation by considering the localness of experts, as revealed through Twitter lists. As mentioned before, the presence of a user in a list is considered as a signal for expertise of that user in that topic.

The intent of this section is to study how the geographical footprint of labelers and experts and the concentration of expertise vary for different topics. For the first part I use two methods. First, I plot heatmaps to provide a visual representation of the distribution of both across the country. Secondly, to quantify the comparison, I calculate the entropy for the distribution of labelers and experts. For comparing the concentration of expertise, I use Gini coefficient.

Through this, I try to address the following questions:

- Does geolocation play a role in expertise/interest in a topic?
- Is the impact of geolocation the same for a topic for expertise and interests?
- Is expertise spread across experts in similar fashion for all topics? How popular are top experts in each topic?

Using the geographical coordinates of the experts and those of the labelers I generated heat maps for the United States for four topics: celebs, food, politics, and tech. Figure 3.4 shows the heat maps reflecting the distribution of experts (the labelees), whereas Figure 3.5 shows the heat maps of the list labelers. It can be easily observed that the geospatial footprint of all topics are not the same. Also, for a topic the spread of labelers and experts also vary distinctly. A few observations:

- The topic `food` has experts from many locations in the country.

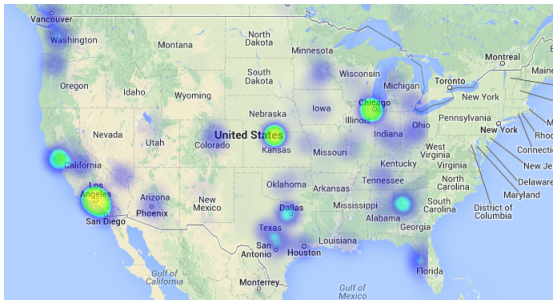
(a) celebs

(b) food

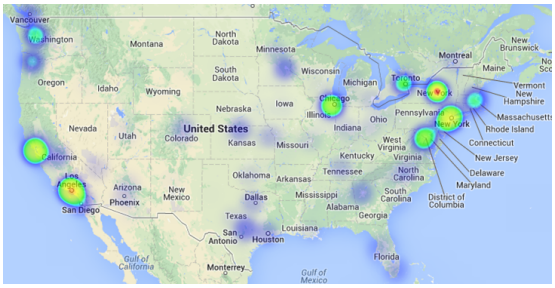
(c) politics

(d) tech

Figure 3.4: Experts heat maps

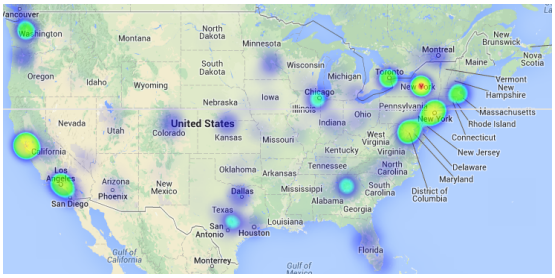


(a) celebs



(b) food

(c) politics



(d) tech

Figure 3.5: Labelers heat maps

- When compared to `food`, the experts from topic `politics` are more concentrated to a few natural home locations in the country – specifically in Washington, D.C. and New York.

- There is a stark contrast in the geospatial footprint of labelers and experts for the `celebs` topic. While the labelers are very well spread out across the length and breadth of the country, the experts are highly concentrated to a few places. This shows that while Twitter users from all over the country are interested in what celebrities tweet, the celebrities are concentrated mainly in Los Angeles and New York. One can observe a similar pattern in `politics` as well.

3.2.1 Localness of Experts: Entropy

To capture the above observations in a quantitative sense, I adopt entropy as a measure of statistical dispersion. For this purpose I use the method of discretizing the earth's surface with a geodesic grid; this allows me to map the latitude, longitude co-ordinates to discrete regions within the United States of America. There are many ways of constructing geodesic grids. Like Serdyukov et al. [40], I use a simple strategy: a grid of square cells of equal degree, such as 1° by 1°. This produces variable-size regions that shrink latitudinally, becoming progressively smaller and more elongated the closer they get towards the poles. Other strategies, such as the quaternary triangular mesh [15], preserve equal area, but are considerably more complex to implement.

Entropy (H). In Kim et. al [30], the authors use entropy as a measure of spatial dispersion of economic activities. The dataset in my work is similar in nature and the model can be directly applied here as well. To illustrate the use of entropy concept here, consider that the area to be considered is divided into k regions in which users, n, are distributed according to n_i; where the subscript i refers to the region i and

the sum of n_i equals total number of users N. The entropy H of users in the country is found using the following equation:

$$H = -\sum_i [p_i * ln\,(p_i)]$$

where, $p_i = n_i/N$. As defined above, H provides a measure of the entropy or dispersion of users n. The value of H ranges from a minimum of 0, if $n_i/N = 1$ and $n_j/N = 0$ for all j not equal to i, to a maximum of $\ln(k)$, if $n_i = n_j$ for all i and j. That is, if all the users are concentrated to a single region, entropy is 0 and it tends towards 1 as the users are spread out across regions. The greater the value of H, the greater the dispersion of users n.

Entropy (F Statistic). Since the entropy H is of little intuitive appeal, it is useful to define the statistic F as follows:

$$F = exp\,(H)$$

The F statistic is a monotonic transformation of H with more intuitive appeal. The F statistic represents the number of equal-sized regions necessary to generate the observed level of entropy or dispersion. The F statistic varies from a minimum of 1, when $H = 0$ and all users n are concentrated in a single region, to a maximum of k, when $H = \ln(\ k)$ and n are uniformly distributed. The F statistic has been used extensively in industrial organization analysis and has been termed the numbers-equivalent of H [25].

Entropy (G Statistic). A similarly useful statistic can be defined as follows:

$$G = H/ln\,(k)$$

Topic	Entropy	F statistic	G statistic
celebs	**3.09**	21.98	0.42
politics	**3.42**	30.57	0.46
tech	3.43	30.88	0.46
finance	3.60	36.60	0.49
food	3.98	53.52	0.54
media	4.10	60.34	0.55
travel	4.11	60.95	0.55
sports	4.31	74.44	0.58

Table 3.4: Dispersion measures for topic experts

Again, G is a monotonic transformation of H with more intuitive appeal. The G statistic represents the relative entropy of users n. The G statistic varies from a minimum of 0, when $H = 0$ and n is concentrated into a single region, to a maximum of 1, and $H = \ln(k)$ when n is uniformly distributed [18].

Entropy of Experts. The values of entropy, F statistic, and G statistic for experts across different topics are listed in Table 3.4. The topics celebs and politics have the least values for entropy. This shows that the experts from these topics are highly concentrated to a few locations. This supports the observation derived from the heatmaps in the previous section. Though celebs is one of the most frequently appearing tag in the lists, all these lists together are made up of experts from select few locations.

Entropy of Labelers. The values of entropy, F statistic, and G statistic for labelers across different topics are listed in the Table 3.5. The values agree with the heat maps for the labelers. The entropy for politics and celebs are high, which is in direct contrast to the experts entropy values. This shows that experts from these topics have a more 'global' effect. Though majority of experts come from select few locations, they are perceived to be experts not just around their geographical neighborhood.

Topic	Entropy	F statistic	G statistic
finance	4.40	81.45	0.59
food	4.43	83.93	0.60
travel	4.72	112.17	0.64
media	4.76	116.75	0.64
tech	4.78	119.10	0.65
politics	4.95	141.17	0.67
celebs	**5.09**	162.39	0.69
sports	**5.10**	164.02	0.69

Table 3.5: Dispersion measures for topic labelers

Their sphere of influence spreads much farther when compared to experts from other topics. This gives a great scope for expert search and recommendation. When users look for queries or experts related to these topics the focus should be more on the topical expertise than on local expertise. In contrast, for food there is a strong local flavor to expertise. There are many experts in many locations having their own local spheres of influence. Though the frequency of labelers for the experts in food are much lower than that of top experts in celebs or politics, it is not necessarily an indicator of the level of expertise. Thus when building systems for expertise recommendation or search on Twitter, the criteria for expertise evaluation should take into account the strong localness signal of food experts.

3.2.2 Concentration of Expertise

On analysis of the data, it was observed that while 24% of all `food` experts appeared more than ten times in the lists labeled `food`, only 16% of `celebs` appeared more than ten times in the `celebs` list. This showed that the expertise in different topics are not dispersed in the same manner. Expertise in the topic `celebs` is more concentrated than in the topic `food`. To understand the contrast, I chose four topics – celebs, tech, food and media and plotted a graph with percentage of experts on the

X axis and cumulative frequency of the lists they appear in on the Y axis. From the graph shown in Figure 3.6, one can easily infer that the expertise in `celebs` is most concentrated; very few experts are contained in most of the `celebs` list created. The top 10% of experts in celebs make up for more than 80% of the `celebs` list while in media it takes close to top 40% of the experts to make up to 80% of the media lists. To further quantify the degree of this type of expertise concentration, I next adopt the Gini coefficient.

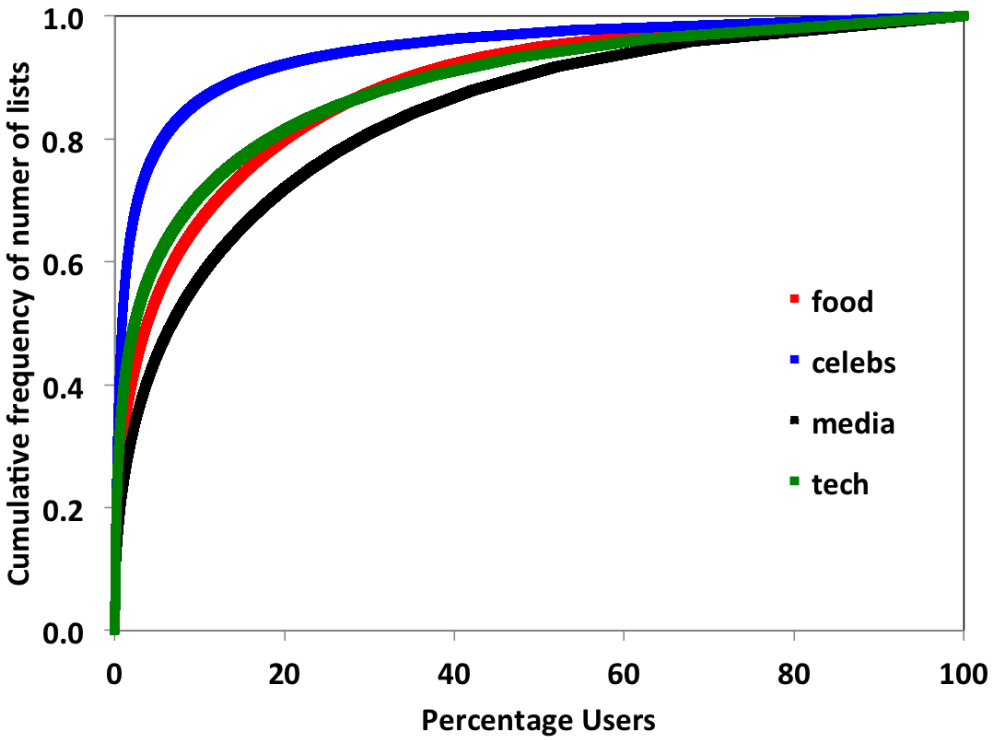


Figure 3.6: Experts vs CDF of lists they appear in

Gini Coefficient. The Gini coefficient is a measure of statistical dispersion intended to represent the income distribution of a nation's residents. The Gini coefficient measures the inequality among values of a frequency distribution. A Gini coefficient

Topic	Gini coefficient
celebs	0.74
tech	0.63
food	0.57
media	0.48

Table 3.6: Gini coefficient values

of zero expresses perfect equality, where all values are the same. A Gini coefficient of one (or 100%) expresses maximal inequality among values. Here Gini coefficient G defined below, is used to measure distribution of expertise, where frequency being the number of times an expert appears in the topic list.

$$X = \sum_{i=1}^{n} \left(\sum_{j=1}^{i} f(j) \right) - \frac{f(i)}{2}$$

$$Y = n * \sum_{i=1}^{n} f(i)$$

$$G = 1 - 2 * \left(\frac{X}{Y} \right)$$

where, n is the number of experts in a topic and $f(i)$ is the number of times the expert i appears in the lists of that topic. A higher Gini coefficient value indicates that expertise is more concentrated i.e., top few experts make up for most of the lists belonging to the topic while a lower value indicates that expertise is more dispersed.

GINI coefficient is a measure to represent the statistical dispersion in the case of non-uniform frequency distribution. The GINI coefficients for the four topics plotted in the graph are shown in Table 3.6. As expected celebs has the highest GINI coefficient while media has the least among the four topics. This throws some light on the distribution of expertise among the topic experts. It is a known fact

that celebrities attract a lot of attention on social media. Overwhelming popularity and attention is mostly limited to a few well known celebrities. The GINI coefficient value goes to show that lists tagged with 'celebs' are dominated by a select set of very popular experts - they make up for a major chunk of the lists. As the value decreases down the table, the inequality in the distribution reduces.

3.2.3 Summary

From the comparative study of the geographical footprints, it can be seen that geolocation does play a role in the expertise/interest of a topic. While topics like food have experts from many locations in the country, topics like politics and celebs have a very small geographical presence concentrated to very few locations. Further, it can be deduced through the heatmaps and entropy values that the geolocation impact on expertise and interest vary greatly. Even locations spread far from the sphere of experts host a sizable number of users interested in the topic. This stark difference is well evident in topics celebs and politics. Lastly, through the use of Gini coefficient, I showed that distribution of expertise is not the same for different topics. There are topics with a few popular experts who make up for a majority of the lists memberships, while in others there are topics where the inequality in the distribution of expertise is not that distinct. From the heatmaps and entropy values, it was evident that the topic food has experts spread out over many locations. The relatively low value of Gini coefficient for food further bolsters the inference of localness of food - expertise is spread out - there are many experts that are locally popular.

3.3 Localness of Topics

In this section, we consider the geographic properties of topics themselves, rather than the labelers or labelees.

Recall that the data contains the geographic coordinates of both the labeler and the labelee in the form of latitudes and longitudes for all the lists. I measure the distance between these two locations using the Haversine distance function, which accounts for the effects of the Earth's spherical shape. For a pair of locations l_1 and l_2, the distance between them is calculated as:

$$\mathcal{D}(l_1, l_2) = 2r \arcsin\left(\sqrt{haversin(\phi_2 - \phi_1) + cos(\phi_1)cos(\phi_2)haversin(\psi_2 - \psi_1)}\right)$$

where, $haversin(\theta) = \sin(\theta/2)^2$ is the Haversine function, D is the distance between the two locations l_1 and l_2, r is the radius of the earth, ϕ_1 and ϕ_2 are the latitude of l_1 and latitude of l_2, and ψ_1 and ψ_2 are the longitude of l_1 and longitude of l_2.

Cheng et. al [6] performed initial analysis on the localness of topics. They tried to understand the geo-spatial properties that were revealed by the lists. For four example topics – `tech`, `entertain`, `travel`, and `food` – the cumulative distribution of frequency of list labeling relationships was plotted over distance. That is, how far apart are the list labelers from the list labelees? The result is shown in Figure 3.7. We can observe that almost 40% of Twitter users who are labelees in a `food` relevant list are within a hundred miles to the labelers. However, only about 10% to 15% of the labelees in a list of other three topics are within a hundred miles to the labelers. In addition, the average distance between a pair of list labeler and list labelee for `food` is also much smaller than the average distance for other topics. These observations

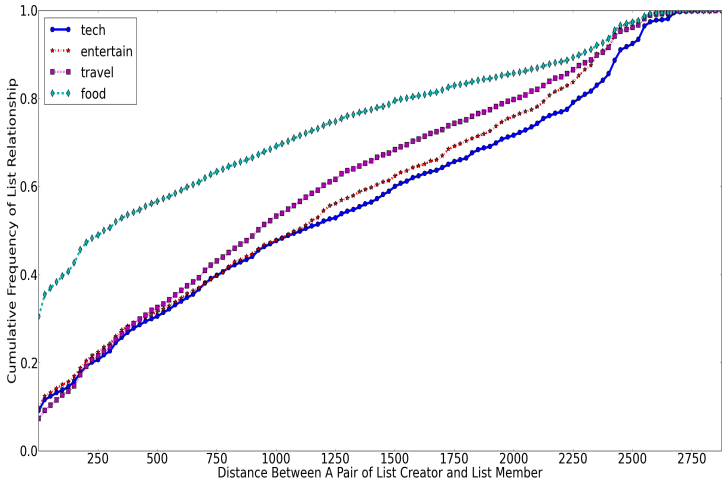


Figure 3.7: Cumulative frequency of list relationship distances

Topic	Entropy
food	3.539
news	3.914
media	3.995
politics	4.080
tech	4.290
celebs	4.360

Table 3.7: Entropy values for list relationship distances

suggest that certain topics are inherently more local than others.

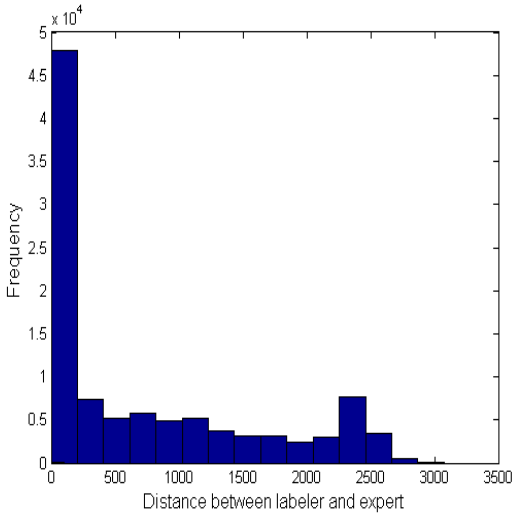
To quantify this comparison, I used entropy on the distance pair values. The entropy thus obtained for various topics are listed in Table 3.7. These values serve as a measure of the dispersion of the labelers with respect to the expert.

The plots in Figure 3.8 show the distances between labeler and labelee for different topics. The topic with lowest entropy value `food` can be seen to have most of the distances concentrated locally. Whereas tech and celebs have distances spread out across the spectrum. Topic celebs with the highest entropy value can be seen to be

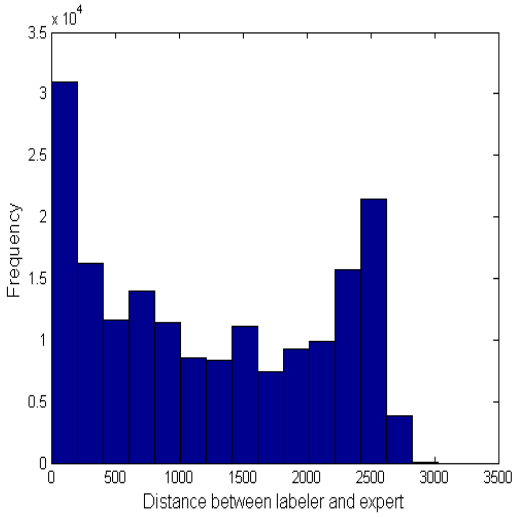
the most dispersed. How does this answer the localness nature of the topics? For food, almost all pairs of distances lie within 250 miles. Since the plot covers all food experts, it represents the localness property of the topic by itself. The radius of spheres of influence for most experts lie within 250 miles (with a few outliers if any). For an expert recommendation/search system, this is a strong signal of localness - users interested in topic food have a tendency to generally care about what the local experts are saying. In contrast, tech is more dispersed in terms of labeler-expert distances. There is a weak localness signal - there are many pairs that lie within the 200 miles range, but there are many pairs that extend up to almost 3000 miles (distance from east coast to west coast across the country). A recommendation system should take into account that users are interested in experts both local and farther from their location. Here is where the topical authority plays an equally important role (if not more) as local authority. Depending upon the geographical location of the user, the two signals need to be judiciously combined to recommend experts that the user would be interested in. Lastly, for celebs, the localness signal is almost negligible. The distance pairs fall almost uniformly across the distance spectrum. Here, the interest for an expert is almost completely dependent on the expert's topical authority with very little emphasis on the geographical proximity to the user's location.

3.3.1 Local Vs Global: Measuring Focus, Entropy, and Spread

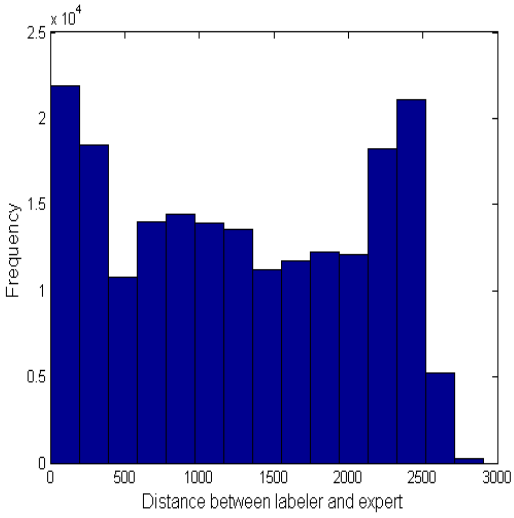
Previous studies of the geographic scope of social media and web resources have typically adopted two types of measures: one considering the intensity of focus and one considering the uniformity of this interest. Here, I adopt three measures (similar to ones for studying hashtags in Kamath et. al [28]): *expert focus* and *expert entropy*, plus a third measure called the *expert spread*.



(a) food



(b) tech



(c) celebs

Figure 3.8: List relationship distances for topics

For every list labeler c who has added an expert to his/her list ($c \in C(e)$, labelers of expert e) and location ($l \in L$) pair, if we let O_l^c be the set of all labelers of e in l, then the probability of a expert's labeler in location l for that expert is defined as:

$$P_l^c = \frac{O_l^c}{\sum_{l \in L}\{O_l^c\}}$$

Expert Focus. Then the *expert focus* for expert e is:

$$\mathcal{F}^e = \max_{l \in L} P_l^c$$

which is simply the maximum probability of observing the labelers at a single location. When an expert's influence isn't local and the expert has labelers from many locations, intuitively the expert's focus will be low, the focus reducing as the labelers are observed at multiple locations. The more local an expert is, presumably the higher the focus will be.

Expert Entropy. The *expert entropy* is defined as:

$$\mathcal{E}^e = -\sum_{l \in L} P_l^c \log_2 P_l^c$$

which measures the randomness in spatial distribution of the labelers of an expert and determines the minimum number of bits required to represent the spread. An expert whose labelers come from only a single location will have an entropy of 0.0. As the labelers spread to more locations, the expert's entropy will increase, reflecting the greater randomness in the distribution.

Expert Spread. While focus and entropy provide insights into an expert's localness, they lack explicit consideration for the distance between the labeler and the expert.

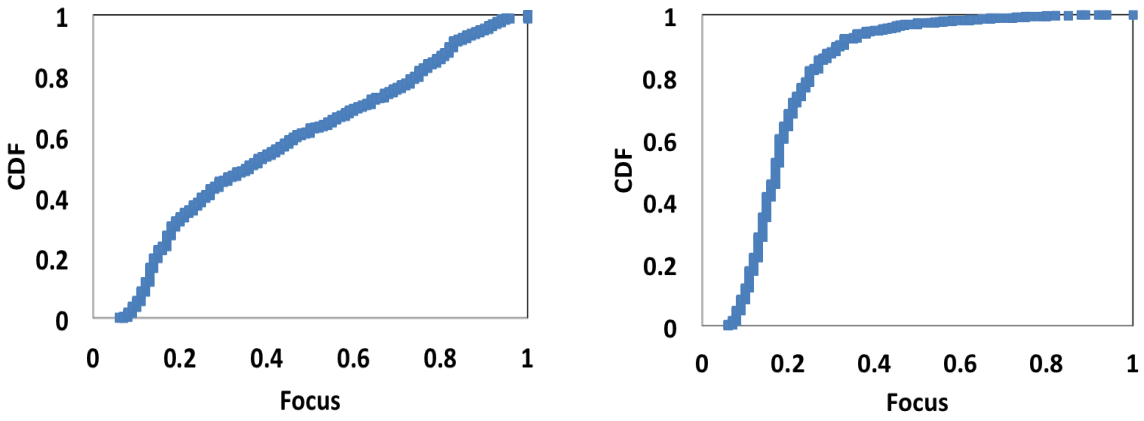
For example, consider two experts – one whose labelers are distributed equally in locations around the expert location, and another one equally distributed between not just local locations but locations farther from the expert location as well. The focus of both experts could be equal and their entropy is 1. Hence, to measure the greater "dispersion" of the second expert's labelers, we define the *expert spread* of expert e as:

$$\mathcal{S}^e = \frac{1}{|C|} \sum_{c \in C} \mathcal{D}(l_c, l_e)$$

which measures the mean distance for all labelers of an expert from the expert's location. Here, l_c is the location of labeler c and l_e is the location of expert e. A local expert with many labelers close to her location will yield a small spread, while a global expert with labelers relatively far from her location will yield a larger spread.

Using these three spatial properties, I analyze the properties of topic localness. I consider the topics `tech` and `food` as examples to compare their geo-characteristics through these measures.

Measuring Topic Focus. Firstly, I consider the focus values for all the `food` experts. For each expert, the focus was calculated using the geo-location information of his/her labelers with the grid concept explained before. The cumulative distribution for focus values for food is shown in Figure 3.9a. We can observe that the distribution is nearly linear, meaning that the focus values for food is almost uniformly distributed. We notice that most labelers are concentrated in one location (high focus). Specifically more than 50% of the experts have a much localized influence i.e., most of their labelers are concentrated to a few locations. In contrast, on observing the graph in Figure 3.9b, that plots the same for `tech` has an initial steep slope and is almost flat for most part after that. More than 80% of the experts have low focus values i.e., their labelers are spread out in many locations. This is indicative of `tech`



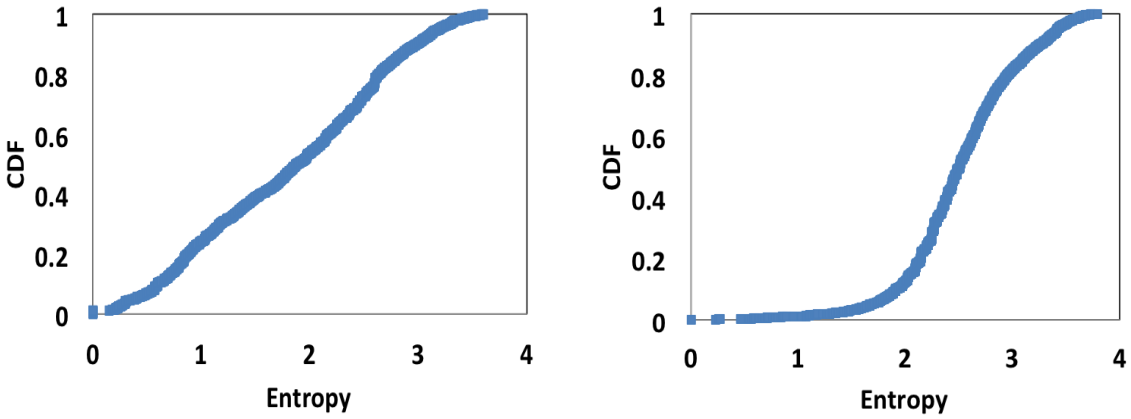
(a) food (b) tech

Figure 3.9: Focus CDF comparison

being a more global topic than `food`.

Measuring Topic Entropy. To further explore this spatial distribution, I next consider the entropy for the two topics. The entropy of zero for an expert indicates that the labelers for that expert in that topic list come from one (2^0) location only, while for example, an entropy value of two indicates that the labelers come almost equally from four (2^2) locations in the grid. The cumulative distribution of entropy of food as shows in Figure 3.10a shows that around 50% of the experts have an entropy lower than 2 i.e., the labelers come from four locations. In contrast only 10% of the tech experts (shown in figure 3.10b) lie within that range. These results show that the majority of food experts have a narrow base of geographic support while tech experts have labeler base spread across the country.

Measuring Topic Spread. While focus and entropy provide insights into a topic's localness, neither directly measures the geographic area over which the labelers are spread out. Using spread definition stated earlier, we plot the spread of the experts in `tech` and `food` lists in Figure 3.11. We observe that most of the food experts



(a) food (b) tech

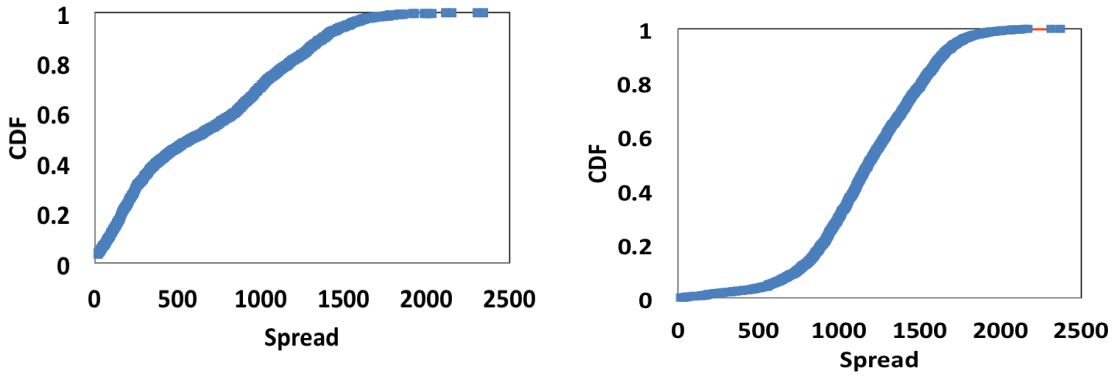
Figure 3.10: Entropy CDF comparison

have a small spread - close to 40% of experts have their labelers within 500 miles of their location. tech experts have farther spread of influence - less than 10% of tech experts have their labelers within 500 miles of their location. Most of the labelers of tech experts have spread values between 1,000 and 2,000 miles.

The analysis I performed on the above two topics was extended to other topics – namely celebs, finance, media, politics. The three measures were calculated in a similar fashion for each of these topics alongside food and tech as well.

1. Focus: For each topic, the focus of each of the experts in that topic were calculated and the CDF was plotted. The comparison of the focus CDF of all the topics were plotted, the result is shown in Figure 3.12a. From the graph, it can be seen that `food` is the most local topic and `celebs` is the least local or most global among all the topics analyzed.

2. Entropy: For each topic, the entropies of each of the experts in that topic were calculated and the CDF was plotted. The comparison of the entropy CDF



(a) food (b) tech

Figure 3.11: Spread CDF comparison

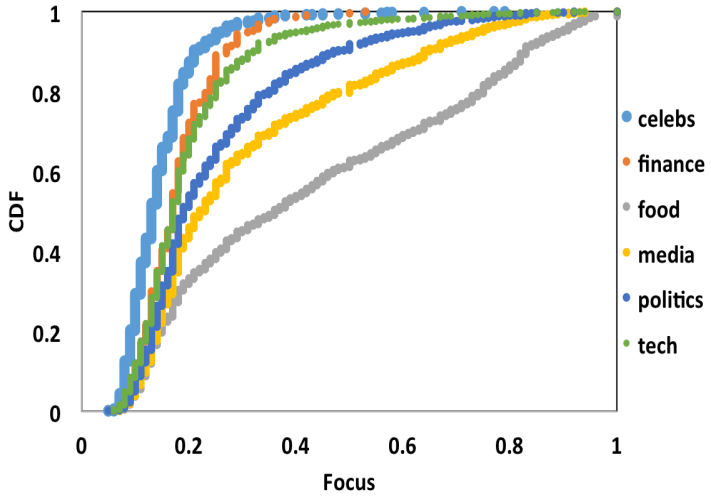
of all the topics are shown in Figure 3.12b. The plot agrees with the above inference.

We can infer the decreasing order of localness for these topics as food, media, politics, tech, finance and celebs. This exactly coincides with the results that we obtained from the entropy values from Table 3.7 using the labeler-labelee distances for the topics.

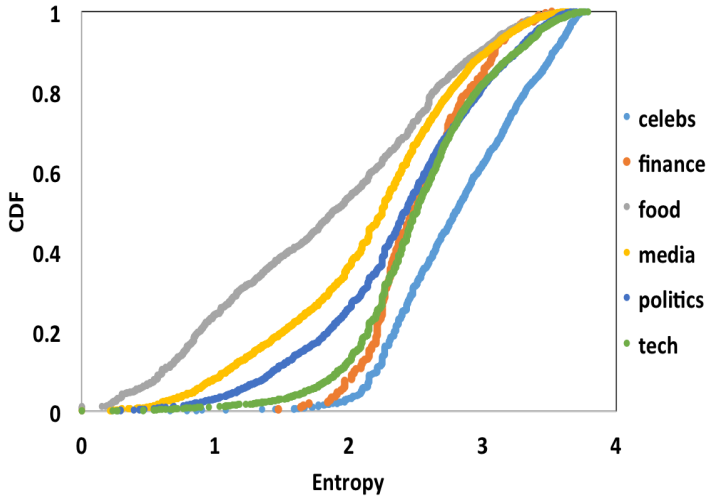
3.3.2 Direct Comparison of Spatial Properties

Through the CDF plots of the three spatial measures, the localness of various topics was studied. Now I turn to directly comparing the focus, entropy and spread values for the topics.

Entropy vs Focus. For each topic, the entropy and focus of the experts were plotted and the result is shown in Figure 3.13. The plot for the topic celebs has a distinctive characteristic - almost all of the experts are concentrated to the top left region - high entropy and low focus. Thus celebs experts have a global impact - the labelers come from many locations and they are dispersed across the country.



(a) food

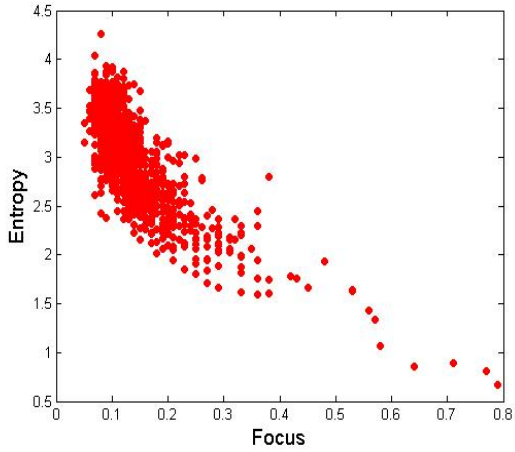


(b) tech

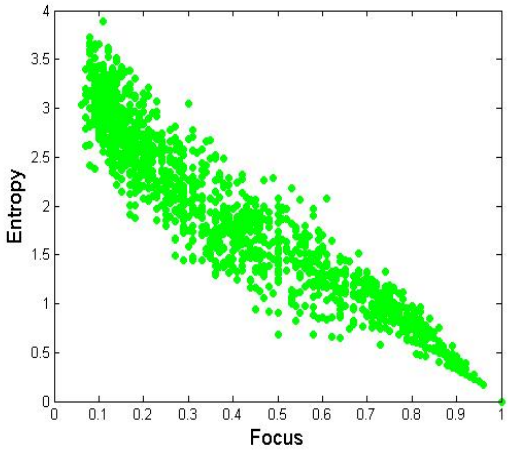
Figure 3.12: Comparison of localness between topics

This observation concurs with the results from previous analysis of the localness of celebs topic. The plot for tech also shows similar properties - most experts fall in the low focus, high entropy group. Similar observations can be made about politics as well. In comparison, for the topic media, the experts are spread out across the spectrum - possibly because of the presence of experts associated with local media and national media - both being equally impactful in many locations. In comparison to all the topics, food has the maximum concentration of experts in the lower right corner - region of high focus and low entropy; which shows that many food experts have labeler presence highly concentrated to a very few locations.

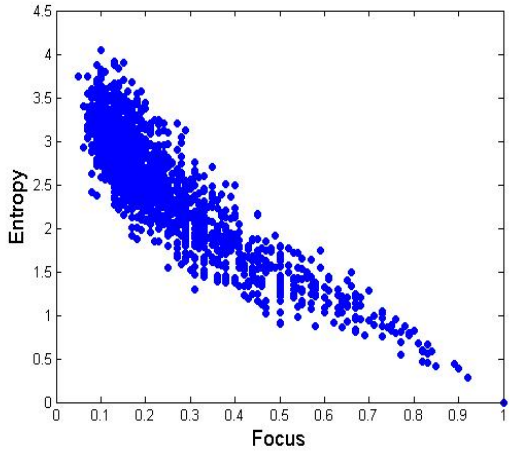
Spread vs Focus. For each topic, the focus and spread were calculated for each expert. For each spread value, the average focus value was plotted and the result is shown in Figure 3.14. As expected, an increasing spread results in a decreasing focus because as the mean labelee-labeler distance increases, the labelers occur in locations of varying distance from the center location of the labeler which in turn reduces the overall focus. As can be observed from the plots, for celebs and tech, the majority of the experts lie in the low focus, large spread region. For food, however, there is a crowd of experts in the high focus region - we observe a steep drop in focus up to 700 miles, followed by a region of almost uniform focus. This initial steep drop of focus indicates that the locations of the labelers are spatially close to the location of the labelee. On a map, the spatial distribution of these points would look like a tight cluster of dots in a small region around the labelee location. The next region where the focus remains almost the same while the spread increases corresponds to labelers who are spatially well distributed but the majority of labelers come from a single location. On a map the spatial distribution for these labelers would have dots spread over a wide region but most of them are concentrated to few locations.



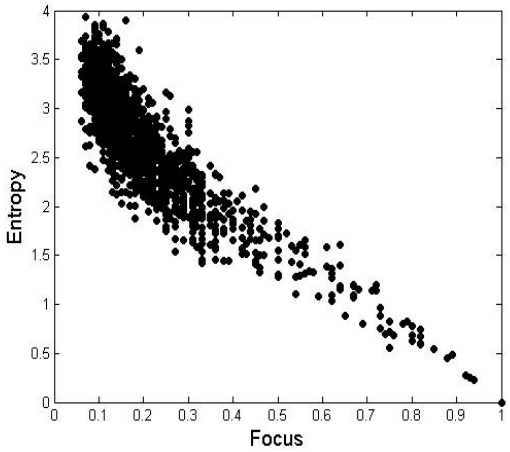
(a) celebs



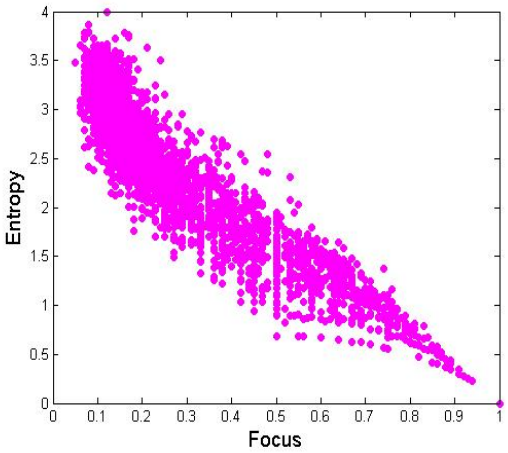
(b) food



(c) politics



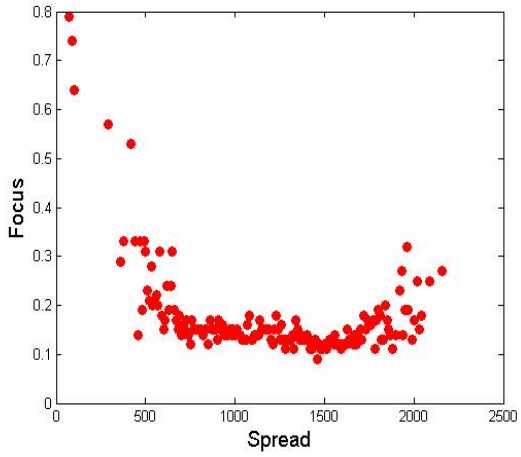
(d) tech



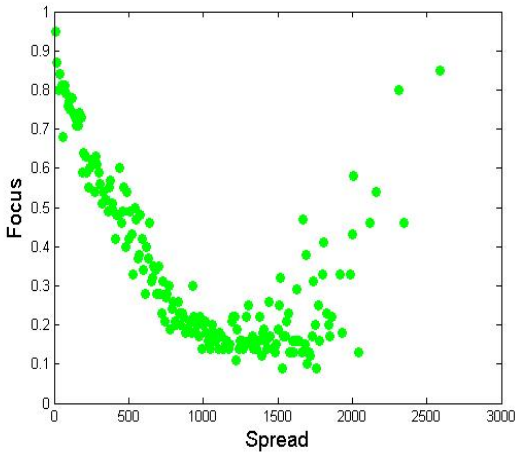
(e) media

Figure 3.13: Entropy vs focus for topics

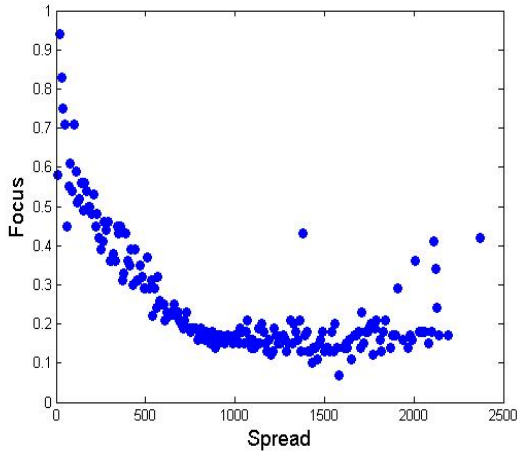
H



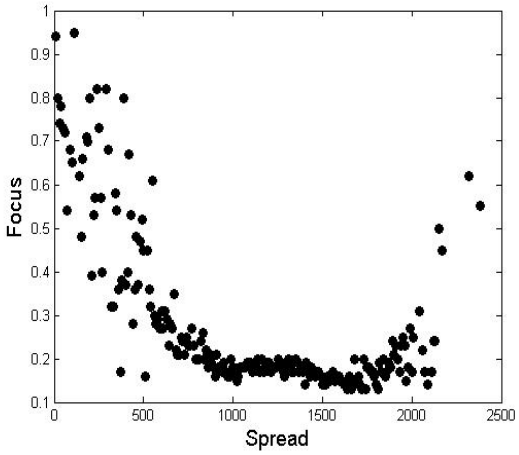
(a) celebs



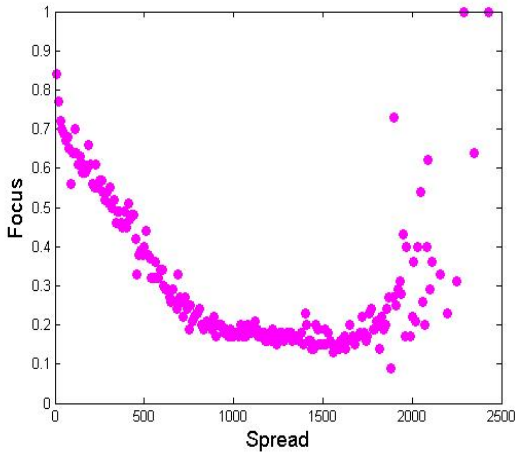
(b) food



(c) politics



(d) tech



(e) media

Figure 3.14: Spread vs focus for topics

Spread vs Entropy. For each topic, the entropy and spread were calculated for each expert. For each spread value, the average entropy value was plotted and the result is shown in Figure 3.15. As expected, an increasing spread results in an increasing entropy. As can be observed from the plots, for celebs and tech, the majority of the experts lie in the high entropy, large spread region. For food, however, there is a crowd of experts in the low entropy region.

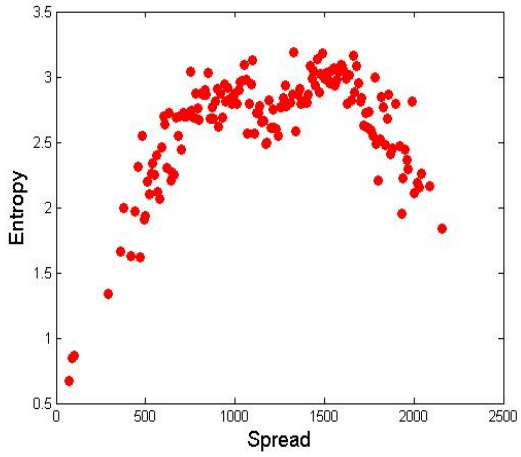
3.4 Localness vs Popularity of experts

How does popularity of an expert impact these measures? We can understand the popularity of an expert in a particular topic to be related to the number of labelers who tag the expert in lists of that topic.

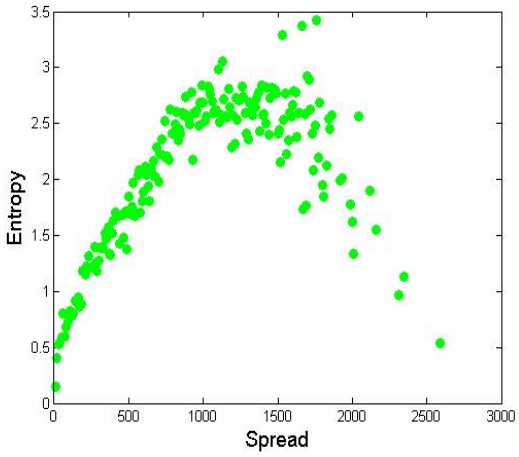
Focus: I performed an analysis on how the expertise of a labelee affects the spatial properties. The plot between focus and count of labelees for experts in topics `food` and `tech` are shown in Figure 3.16. The plot for tech indicates a clear trend; experts with many labelers have very low focus values. From this it can be inferred that for the most popular among the tech experts the labelers are diffused throughout the geographical area while for experts with less than 20 labelers, these labelers have high geographic concentration. The plot for food does not strictly fit this pattern. Given the local nature of the topic food, many experts have high focus values. From the plot we can observe that there are popular experts who have higher than average focus values - this emphasizes the localness nature of food compared to that of tech. While in tech, most popular experts have labelers from a wide spread of geographical locations, quite a few popular food experts have their labelers concentrated to local locations.

Entropy: Figure 3.17 shows the plot between entropy and count of labelees for experts in topics `food` and `tech`. Again, the plot for tech shows a clear trend - popular experts

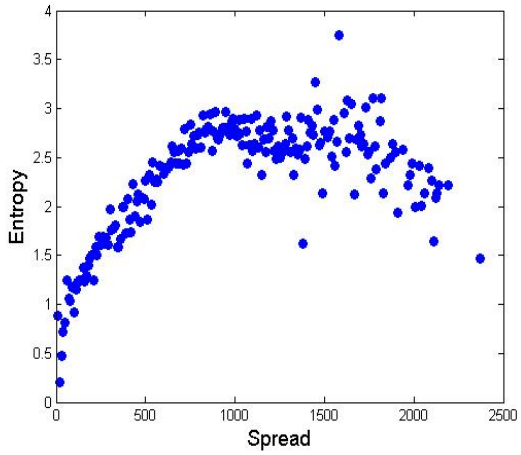
H



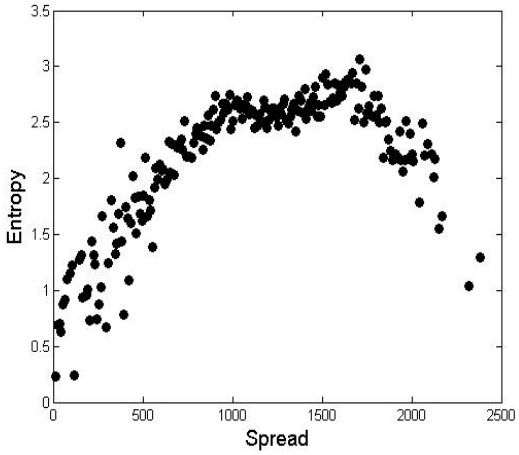
(a) celebs



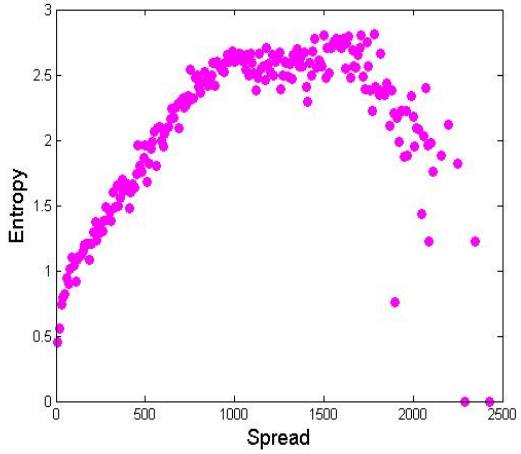
(b) food



(c) politics

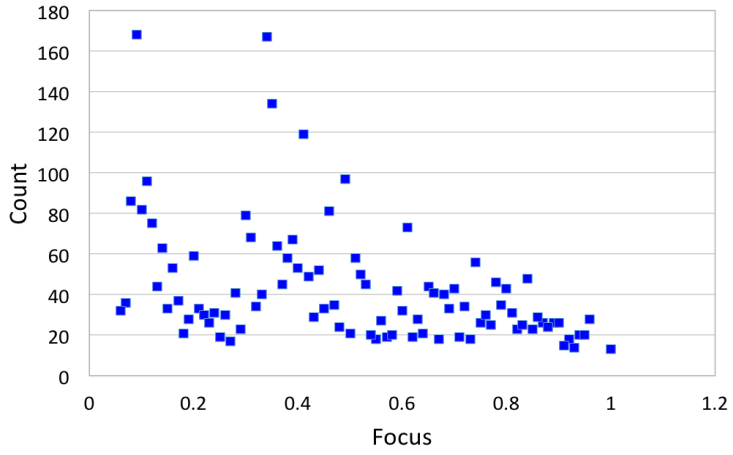


(d) tech

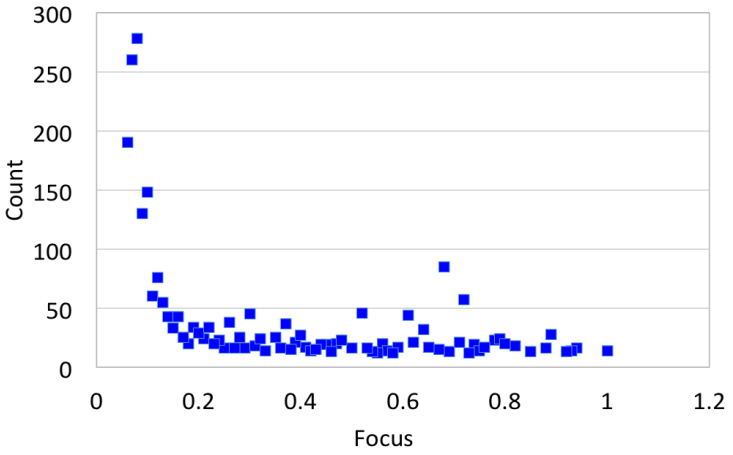


(e) media

Figure 3.15: Spread vs entropy for topics

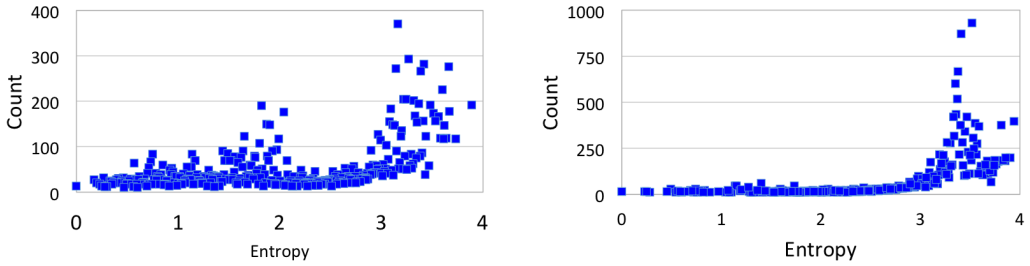


(a) food



(b) tech

Figure 3.16: Focus vs frequency of labelers



(a) food (b) tech

Figure 3.17: Entropy vs frequency of labelers

have higher values of entropy indicating that the distribution of their labelees is more diffused while labelers with lesser expertise have a more uniform distribution of labelees. Though in general the entropy plot for food agrees with the above inference, there are a few outliers. Few experts with many labelers have a low value of entropy pointing towards the local nature of the topic. Labelees with high level expertise need not necessarily attract labelers from across the geographical spectrum, they could have labelers who are uniformly distributed in a few local locations.

Spread: Figure 3.18 shows the plot between spread and count of labelees for experts in topics `food` and `tech`. In tech, we can observe that most of the popular experts have spread values between 1,000 to 2,000 miles - their sphere of influence stretches out much farther than the less popular experts. Again in the case of `food`, the rule doesn't strictly apply. There are many popular local experts - whose labelers are at an average distance within 1,000 miles. Spread is a good indicator of localness of the topic since it is a measure of the labeler-labelee distance. On comparing the two graphs, the localness of food with respect to tech is clearly brought out. There are a considerable number of food experts who have more than 40 labelers whose spread is within 900 miles. On the same scale however, tech has almost no expert in that

range.

From the above plots, we can observe in general that popular experts have a larger sphere of influence and their labelers are more diffused geographically. But food shows anamolous behavior - few experts have many labelers and still have a small sphere of influence. These plots are a further proof to the local nature of the topic food.

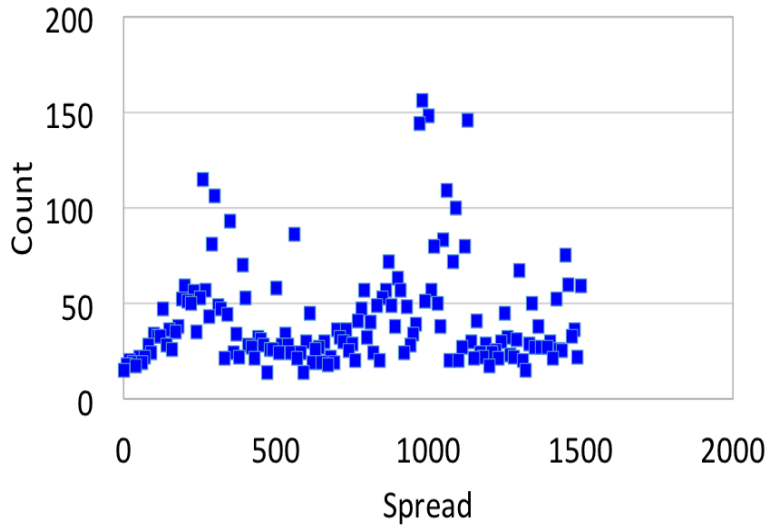
3.5 How Does Topic Localness Vary Across Locations?

In the above section, there was a detailed analysis on the geo-spatial impact on topic expertise. Using different metrics, it was inferred that expertise in topic `food` is localized in comparison to other topics that were considered. For spatial analysis, I used labeler locations from all over the United States for lists labeled `food`. food experts come from different parts of the country, the spheres of influence of experts aren't the same everywhere. To study how the topic food varies in expertise across different geographical areas of the country, I considered three different cities - New York, Chicago and San Francisco.

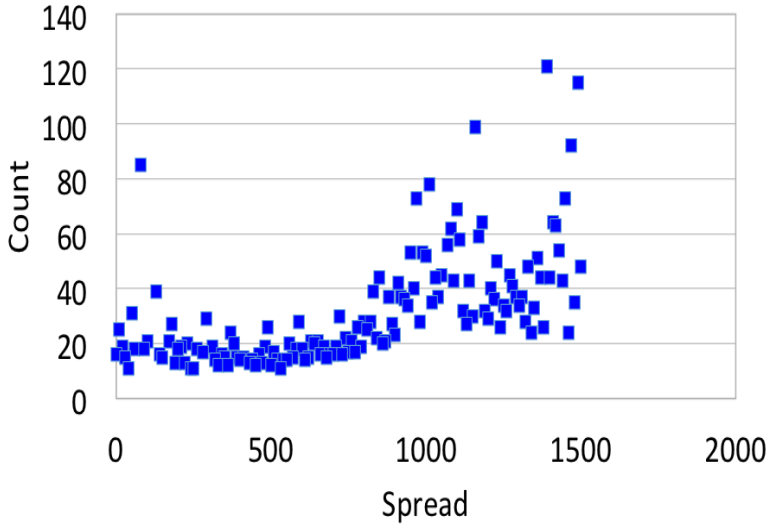
3.5.1 Spatial Measures

Firstly, the experts from these cities were identified using their geo-location information. Thereafter, the geolocation of the set of labelers who have these experts in their lists labeled `food` were used to calculate the three measures discussed earlier.

Focus. Figure 3.19a shows the comparison in the focus CDF plots of the three cities for the topic food. The CDF curve for New York is quite steep, while the curves for SF and Chicago are quite similar and have lesser slopes. Only around 30% of the experts from SF and Chicago lie in the low focus region, while more than 75% of the experts from New York lie within that range. The labelers who add the experts from SF and Chicago in their lists are mainly concentrated in a single location.



(a) food



(b) tech

Figure 3.18: Spread vs frequency of labelers

Entropy. Figure 3.19b shows the comparison in the entropy CDF plots of the three cities. The curve trend is exactly complimentary to that of focus plot - most of the experts from SF and Chicago have lower entropies and majority of New York experts have high values of entropy. Between SF and Chicago, the focus curve is slightly steeper for SF and vice versa for the entropy curve. This is a direct indication to the labelers of experts from SF being more dispersed and diffused than that of experts from Chicago.

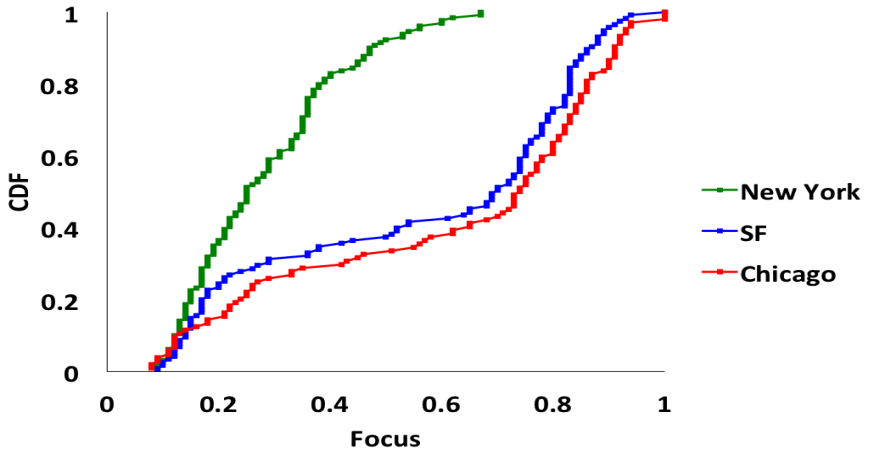
Spread. Figure 3.19c shows the comparison in the spread CDF plots for the three cities for food. From the plot, it can be observed that while only half of the experts from New York have their labelers with spread of within 750 miles, 60% of experts from SF and 80% of experts from Chicago lie within that range. In fact, the spread of all the experts from Chicago are within 1000 miles.It indicates that the sphere of influence of experts from Chicago is smaller when compared to that of experts from SF and New York. The food experts from Chicago have a localized influence, while experts from New York and San Francisco attract labelers not only locally but also from other locations that are farther from their location.

3.5.2 Heat Maps

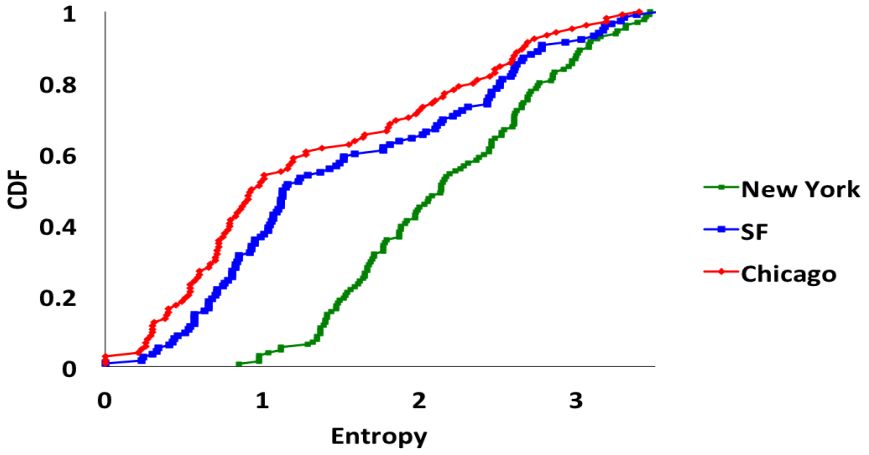
To evaluate the validity of the above inference, I plotted the relevant geo-locations on the map of the United States of America using Google Maps API. For each of the three cities, I collected the latitude, longitude of all the labelers in the country who have added experts from the city into their food lists. The result is shown in the three maps in Figure 3.20. From the first map that is of labelers for experts from Chicago, it can be observed that almost all the labelers are from in and around the city of Chicago. There isn't any trace of labelers for Chicago experts outside of their localized sphere of influence. A very similar trend can be observed in the second

map that shows the heat map for the labelers for SF experts. Most of the labelers come from SF and its nearby locations, but there is a faint trace of labelers from other parts of the state as well, notably Los Angeles. An interesting point here is the small influence of SF experts in parts of New York. This explains the anomaly in the spread CDF curve for SF experts. Since SF and NY are very far apart, the average spread for a few experts who have labelers in NY as shown in the heat map, are biased to a higher value and hence the difference in curve in comparison with NY above 1200 miles. The last figure shows the heat map for NY experts' labelers. Again, we see a high concentration of labelers from local regions. But unlike in the other two cities, here we can see labeler presence in other parts of the country as well. There is a a noticeable trace of labeler population in parts of California, Boston, D.C. and Chicago. The measures we defined above for this purpose capture the spatial properties of the impact and influence of the food experts located in the three cities.

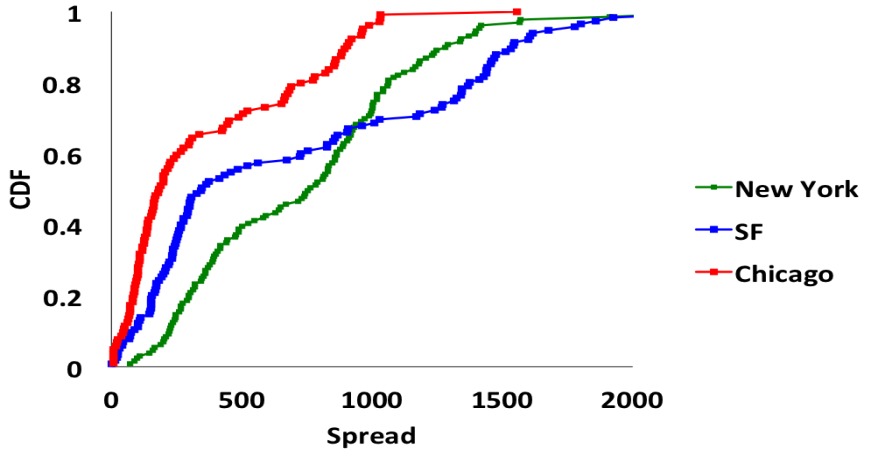
From the above analysis it can be seen that the spatial properties of experts belonging to the same topic can also vary between regions. The topic 'food' is popular in New York and the experts attract labelers from larger geographical spectrum and the labelers are diffused across these locations. This is a rough comparison of the properties as the cities vary significantly in their surface area, population density and also the number of active twitter users. To better compare the results, a more complicated approach to normalize the data observed is required.



(a) Focus

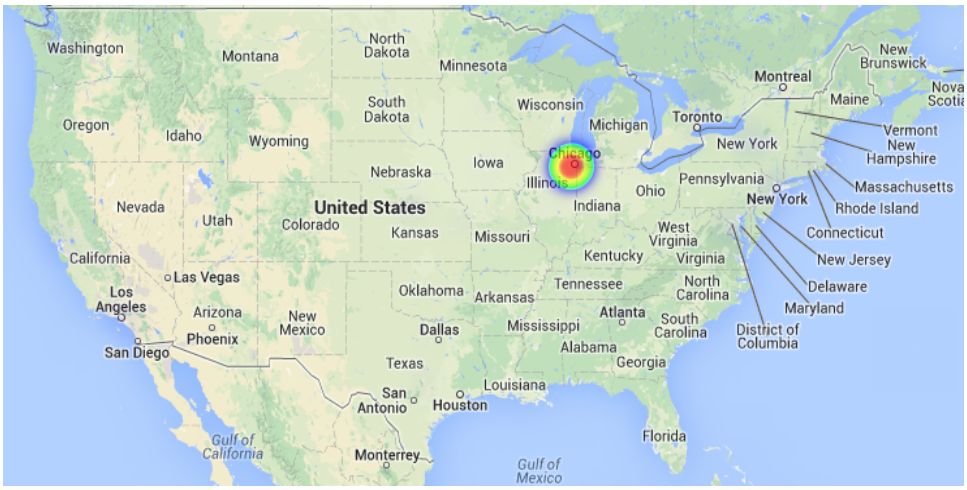


(b) Entropy



(c) Spread

Figure 3.19: Spatial measures for food



(a) Chicago

(b) San Francisco

(c) New York

Figure 3.20: Heat maps for food experts by cities

4. CONCLUSION AND FUTURE WORK

With the advent of GPS-enabled smartphones, the impact of geolocation in social media interaction has increased exponentially. This work performs a detailed analysis on the localness of topic expertise. I studied extensively the effect of geolocation on expertise and employed statistical measures to compare and contrast the localness effect of different topic experts. This study can be used as a comprehensive tool for building recommendation systems on Twitter, among other new social and geo-aware applications.

The system suffers from the limitation arising out of the use of grid-based method for identifying locations. This method does not take into account the variance in area, density of population and population of twitter users in different grids. This can be improved upon by using other techniques to map locations in an efficient manner considering all these features as well. Secondly, I use a simplified method for topic extraction. For large, real world data, there needs to be a more robust way to extract and process tags. This work uses Twitter lists as the sole source for expert information. Like any other crowdsourced method, this system is susceptible to spamming. Since expertise is simply interpreted as the number of lists an user appears in, it is difficult to filter out the genuine signals from the malicious ones. One way to counter this would be to use sophisticated link-based models to capture expertness. Also, list information can be augmented by user bio, tweets content and follower information to make the expert system more reliable.

As part of future work, I would like to extend the analysis to a global level and contrast the topic study between different countries of the world. In addition, I would like to build on this analysis and use the results to develop a model for

capturing localness of topic expertise on Twitter and other social media. Some of the possible implementations are list based expert recommendation system, expert search on twitter that encompasses both topic expertise and local expertise and predicting label relationship between user and expert based on the location of both and the label tag.

REFERENCES

[1] Judd Antin, Marco de Sa, and Elizabeth F. Churchill. Local experts and online review sites. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion*, CSCW '12, pages 55–58. ACM, 2012.

[2] Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. Spatial variation in search engine queries. In *Proceedings of the 17th international conference on World Wide Web*, pages 357–366. ACM, 2008.

[3] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

[4] Michael Bernstein, Desney Tan, Greg Smith, Mary Czerwinski, and Eric Horvitz. Collabio: a game for annotating people within social networks. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, pages 97–100. ACM, 2009.

[5] Anders Brodersen, Salvatore Scellato, and Mirjam Wattenhofer. Youtube around the world: geographic popularity of videos. In *Proceedings of the 21st international conference on World Wide Web*, pages 241–250. ACM, 2012.

[6] Zhiyuan Cheng, James Caverlee, Himanshu Barthwal, and Vandana Bachani. Finding local experts on twitter. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 241–242. International World Wide Web Conferences Steering Committee, 2014.

[7] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the*

19th ACM international conference on Information and knowledge management, pages 759–768. ACM, 2010.

[8] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z Sui. Exploring millions of footprints in location sharing services. *ICWSM*, 2011:81–88, 2011.

[9] Jeremy W. Crampton, Mark Graham, Ate Poorthuis, Taylor Shelton, Monica Stephens, Matthew W. Wilson, and Matthew Zook. Beyond the geotag: situating big data and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, 40(2):130–139, 2013.

[10] Justin Cranshaw, Raz Schwartz, Jason I. Hong, and Norman M. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, editors, *ICWSM*. The AAAI Press, 2012.

[11] Arie Croitoru, Andrew Crooks, Jacek Radzikowski, and Anthony Stefanidis. Geosocial gauge: a system prototype for knowledge discovery from social media. *International Journal of Geographical Information Science*, 27(12):2483–2508, 2013.

[12] Nilesh Dalvi, Ravi Kumar, and Bo Pang. Object matching in tweets with spatial models. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 43–52. ACM, 2012.

[13] Clodoveu A Davis Jr, Gisele L Pappa, Diogo Rennó Rocha de Oliveira, and Filipe de L Arcanjo. Inferring the location of twitter messages based on user relationships. *Transactions in GIS*, 15(6):735–751, 2011.

[14] Junyan Ding, Luis Gravano, and Narayanan Shivakumar. Computing geographical scopes of web resources. 2000.

[15] Geoffrey Dutton. Encoding and handling geospatial data with hierarchical triangular meshes. In *Proceeding of 7th International symposium on spatial data handling*, volume 43. Netherlands: Talor & Francis, 1996.

[16] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010.

[17] Laura Ferrari, Alberto Rosi, Marco Mamei, and Franco Zambonelli. Extracting urban patterns from location-based social networks. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 9–16. ACM, 2011.

[18] Charles B Garrison and Albert S Paulson. An entropy measure of the geographic concentration of economic activity. *Economic Geography*, pages 319–324, 1973.

[19] Judith Gelernter and Nikolai Mushegian. Geo-parsing messages from microtext. *Transactions in GIS*, 15(6):753–773, 2011.

[20] Debarchana Ghosh and Rajarshi Guha. What are we tweetingabout obesity? mapping tweets with topic modeling and geographic information system. *Cartography and Geographic Information Science*, 40(2):90–102, 2013.

[21] Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 575–590. ACM, 2012.

[22] Ido Guy, Uri Avraham, David Carmel, Sigalit Ur, Michal Jacovi, and Inbal Ronen. Mining expertise and interests from social media. In *Proceedings of the*

22Nd International Conference on World Wide Web, WWW '13, pages 515–526, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[23] Scott A. Hale, Devin Gaffney, and Mark Graham. Where in the world are you? geolocation and language identification in twitter. *Proceedings of ICWSM12*, pages 518–521, 2012.

[24] Damon Horowitz and Sepandar D Kamvar. The anatomy of a large-scale social search engine. In *Proceedings of the 19th international conference on World wide web*, pages 431–440. ACM, 2010.

[25] Ira Horowitz. Numbers-equivalents in us manufacturing industries: 1954, 1958 and 1963. *Southern Economic Journal*, pages 396–408, 1971.

[26] Jonny Huck, Duncan Whyatt, and Paul Coulton. Challenges in geocoding socially-generated data. 2012.

[27] Krishna Y Kamath, James Caverlee, Zhiyuan Cheng, and Daniel Z Sui. Spatial influence vs. community influence: modeling the global spread of social media. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 962–971. ACM, 2012.

[28] Krishna Y Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In *Proceedings of the 22nd international conference on World Wide Web*, pages 667–678. International World Wide Web Conferences Steering Committee, 2013.

[29] Dongwoo Kim, Yohan Jo, Il-Chul Moon, and Alice Oh. Analysis of twitter lists as a potential source for discovering latent characteristics of users. In *ACM CHI Workshop on Microblogging*, 2010.

[30] Tschangho John Kim and Gerrit Knaap. The spatial dispersion of economic activities and development trends in china: 1952–1985. *The Annals of Regional Science*, 35(1):39–57, 2001.

[31] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[32] Linna Li, Michael F. Goodchild, and Bo Xu. Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. *Cartography and Geographic Information Science*, 40(2):61–77, 2013.

[33] Bruce H Mayhew and Roger L Levinger. Size and the density of interaction in human aggregates. *American Journal of Sociology*, pages 86–110, 1976.

[34] David W. McDonald and Mark S. Ackerman. Expertise recommender: A flexible recommendation system and architecture. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, pages 231–240, New York, NY, USA, 2000. ACM.

[35] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare. *ICWSM*, 11:70–573, 2011.

[36] Aditya Pal and Scott Counts. Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 45–54. ACM, 2011.

[37] Adam Sadilek, Henry Kautz, and Jeffrey P Bigham. Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 723–732. ACM, 2012.

[38] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, 2010.

[39] Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo. Socio-spatial properties of online location-based social networks. *ICWSM*, 11:329–336, 2011.

[40] Pavel Serdyukov, Vanessa Murdock, and Roelof Van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491. ACM, 2009.

[41] Blake Shaw, Jon Shea, Sidhartha Sinha, and Andrew Hogue. Learning to rank for spatiotemporal search. In *WSDM*, 2013.

[42] Anthony Stefanidis, Amy Cotnoir, Arie Croitoru, Andrew Crooks, Matthew Rice, and Jacek Radzikowski. Demarcating new boundaries: mapping virtual polycentric communities through social media content. *Cartography and Geographic Information Science*, 40(2):116–129, 2013.

[43] Hao Wang, Manolis Terrovitis, and Nikos Mamoulis. Location recommendation in location-based social networks using user check-in data. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 364–373. ACM, 2013.

[44] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.

[45] Chen Xu, David W. Wong, and Chaowei Yang. Evaluating the geographical awareness of individuals: an exploratory analysis of twitter data. *Cartography and Geographic Information Science*, 40(2):103–115, 2013.

[46] Yuto Yamaguchi, Toshiyuki Amagasa, and Hiroyuki Kitagawa. Tag-based user topic discovery using twitter lists. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 13–20. IEEE, 2011.